Inference for Multiple Linear Regression

Nate Wells

Math 141, 4/20/22

Outline

In this lecture, we will. . .

- Review framework for multilinear regression
- Discuss inference procedures for MLR models
- Investigate tools for "Model Selection"

Section 1

Multiple Linear Regression

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- We use the following R code to fit and summarize a linear model:

```r
mod<-lm(Y ~ X1 + X2 + X3, data = my_data)
get_regression_table(mod)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     3.26     7.94       0.41   0.686   -13.3    19.8
## 2 X1           -1.24     0.313     -3.95   0.001    -1.89   -0.584
## 3 X2            2.68     1.94       1.38   0.182    -1.36    6.72
## 4 X3            3.20     0.397      8.06   0         2.37    4.02
```

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- We use the following R code to fit and summarize a linear model:

```
mod<-lm(Y ~ X1 + X2 + X3, data = my_data)
get_regression_table(mod)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     3.26     7.94       0.41   0.686   -13.3     19.8
## 2 X1           -1.24     0.313     -3.95   0.001    -1.89    -0.584
## 3 X2            2.68     1.94       1.38   0.182    -1.36     6.72
## 4 X3            3.20     0.397      8.06   0         2.37     4.02
```

- Which gives us our linear regression formula:

$$\hat{Y} = 3.26 - 1.24 \cdot X_1 + 2.68 \cdot X_2 + 3.2 \cdot X_3$$

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- We use the following R code to fit and summarize a linear model:

```
mod<-lm(Y ~ X1 + X2 + X3, data = my_data)
get_regression_table(mod)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     3.26     7.94       0.41   0.686  -13.3     19.8
## 2 X1           -1.24     0.313     -3.95   0.001   -1.89    -0.584
## 3 X2            2.68     1.94       1.38   0.182   -1.36     6.72
## 4 X3            3.20     0.397      8.06   0         2.37     4.02
```

- Which gives us our linear regression formula:

$$\hat{Y} = 3.26 - 1.24 \cdot X_1 + 2.68 \cdot X_2 + 3.2 \cdot X_3$$

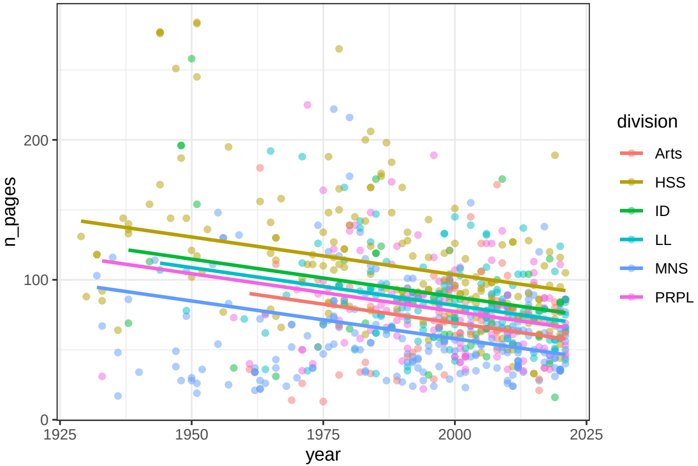
- The slope on each variable indicates the changed in the predicted value of Y per unit change in that variable, **with all other variables held constant**

Reed Theses

- How does page count vary across year of publication **and** division?

Reed Theses

- How does page count vary across year of publication **and** division?



Reed Theses

- How does page count vary across year of publication **and** division?

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

$$\text{Pages} = 1149 - 0.54 \cdot \text{Year} + 34.5 \cdot \text{HSS} + 18.7 \cdot \text{ID} + 12.6 \cdot \text{LL} - 11.2 \cdot \text{MNS} + 8.3 \cdot \text{PRPL}$$

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0      902.     1396.
## 2 year             -0.54     0.063     -8.58   0       -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0       23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011    4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031    1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042  -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162   -3.34    19.9
```

$$\text{Pages} = 1149 - 0.54 \cdot \text{Year} + 34.5 \cdot \text{HSS} + 18.7 \cdot \text{ID} + 12.6 \cdot \text{LL} - 11.2 \cdot \text{MNS} + 8.3 \cdot \text{PRPL}$$

- Which division is used as the baseline? What does the intercept represent?

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

$$\text{Pages} = 1149 - 0.54 \cdot \text{Year} + 34.5 \cdot \text{HSS} + 18.7 \cdot \text{ID} + 12.6 \cdot \text{LL} - 11.2 \cdot \text{MNS} + 8.3 \cdot \text{PRPL}$$

- Which division is used as the baseline? What does the intercept represent?

- What does the coefficient of 34.5 on HSS mean in context?

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

$$\text{Pages} = 1149 - 0.54 \cdot \text{Year} + 34.5 \cdot \text{HSS} + 18.7 \cdot \text{ID} + 12.6 \cdot \text{LL} - 11.2 \cdot \text{MNS} + 8.3 \cdot \text{PRPL}$$

- Which division is used as the baseline? What does the intercept represent?
- What does the coefficient of 34.5 on HSS mean in context?
- What does the coefficient of -0.54 on Year mean in context?

Hypothesis Testing

- The regression table provides p-values for each slope in the model.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
    - But what hypotheses are being tested?

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
    - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

  - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

  - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

- Reminder: The p-value is the probability of obtaining a statistic as extreme as the observed statistic, **if the null hypothesis were true**.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
    - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
    - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

    - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

- Reminder: The p-value is the probability of obtaining a statistic as extreme as the observed statistic, **if the null hypothesis were true**.
- The standard error, statistic, and p-values are all calculated using theory-based methods.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

  - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

- Reminder: The p-value is the probability of obtaining a statistic as extreme as the observed statistic, **if the null hypothesis were true**.
- The standard error, statistic, and p-values are all calculated using theory-based methods.
  - But the formula is very complicated, requiring matrices and linear algebra (If interested, take Math 392)

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

- For which hypothesis tests would you reject H_0?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0        902.    1396.
## 2 year             -0.54     0.063     -8.58   0         -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0         23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011      4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031      1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042    -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162     -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?

- For which would you fail to reject H_0?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?
- For which would you fail to reject H_0?
  - What does this mean in context?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     1149.     126.          9.14   0      902.    1396.
## 2 year            -0.54     0.063      -8.58   0        -0.664   -0.416
## 3 divisionHSS     34.5      5.61        6.16   0        23.5     45.5
## 4 divisionID      18.7      7.38        2.53   0.011     4.22    33.2
## 5 divisionLL      12.6      5.84        2.16   0.031     1.16    24.1
## 6 divisionMNS    -11.2      5.50       -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL     8.28     5.92        1.40   0.162    -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?
- For which would you fail to reject H_0?
  - What does this mean in context?
- What is the relationship between the `estimate`, the `std_error` and `lower_ci`, `upper_ci`?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term         estimate std_error statistic p_value lower_ci upper_ci
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     1149.     126.         9.14   0       902.    1396.
## 2 year            -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS    34.5       5.61       6.16   0        23.5     45.5
## 4 divisionID     18.7       7.38       2.53   0.011     4.22    33.2
## 5 divisionLL     12.6       5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS   -11.2       5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL    8.28      5.92       1.40   0.162    -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?
- For which would you fail to reject H_0?
  - What does this mean in context?
- What is the relationship between the `estimate`, the `std_error` and `lower_ci`, `upper_ci`?
- How does the coefficient on `year` in the MLR model compare to the coefficient on `year` in the SLR model?

Section 2

Model Assumptions for MLR

Model Assumptions

- In order to responsibly use MLR to make inference, we need. . .

Model Assumptions

- In order to responsibly use MLR to make inference, we need...

1. The relationship between explanatory and response variables must be approximately multilinear linear. (**Linear**)
2. The observations should be independent of one another. (**Independence**)
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)

Model Assumptions

- In order to responsibly use MLR to make inference, we need...

1. The relationship between explanatory and response variables must be approximately multilinear linear. (**Linear**)
2. The observations should be independent of one another. (**Independence**)
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)

- How do we check some of these conditions? Why can't we create a scatterplot of residuals as we did for SLR?

Model Assumptions

- In order to responsibly use MLR to make inference, we need...

1. The relationship between explanatory and response variables must be approximately multilinear linear. (**Linear**)
2. The observations should be independent of one another. (**Independence**)
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)

- How do we check some of these conditions? Why can't we create a scatterplot of residuals as we did for SLR?
    - Instead, we will use a scatterplot of residuals vs **predicted values**

Residuals vs Fitted Values

```
mod_res <- get_regression_points(theses_mlr)
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

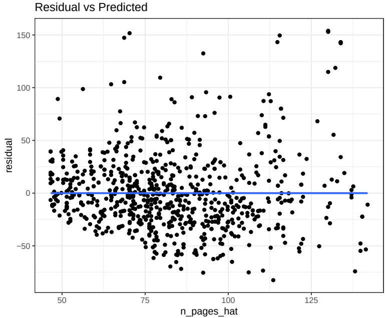
```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

Residuals vs Fitted Values

```
mod_res <- get_regression_points(theses_mlr
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

```
mod_res %>% ggplot(aes(x = n_pages_hat, y = residual)
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Residual vs Predicted")
```

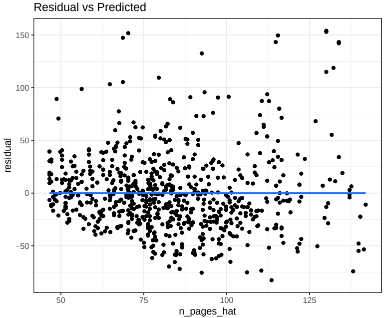


Residuals vs Fitted Values

```
mod_res <- get_regression_points(theses_mlr
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

```
mod_res %>% ggplot(aes(x = n_pages_hat, y = residual)
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Residual vs Predicted")
```



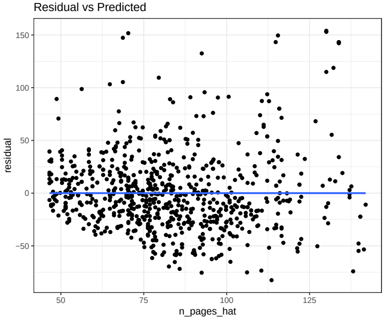
- When analyzing residual vs. predicted plots, look for. . .

Residuals vs Fitted Values

```r
mod_res <- get_regression_points(theses_mlr
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

```r
mod_res %>% ggplot(aes(x = n_pages_hat, y = residual))
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Residual vs Predicted")
```



- When analyzing residual vs. predicted plots, look for...
  - Non-linear patterns
  - Increasing variability across range of predicted values
  - Outliers with atypical predicted value or large residual

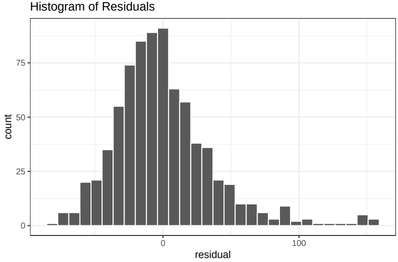
Distribution of Residuals

- We can still look at the histogram and QQ Plot of residuals, as we did for SLR:

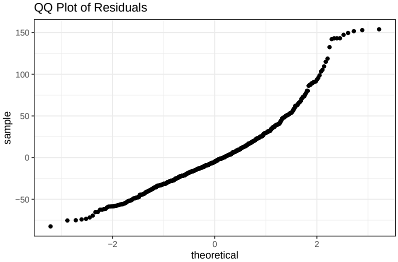
Distribution of Residuals

- We can still look at the histogram and QQ Plot of residuals, as we did for SLR:

```
ggplot(mod_res, aes(x = residual))+
  geom_histogram(bins = 30, color = "white")+
  labs(title = "Histogram of Residuals")
```



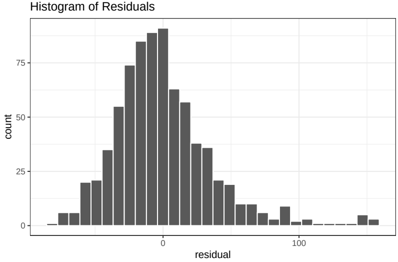

```
ggplot(mod_res, aes(sample = residual))+
  geom_point(stat = "qq")+
  labs(title = "QQ Plot of Residuals")
```



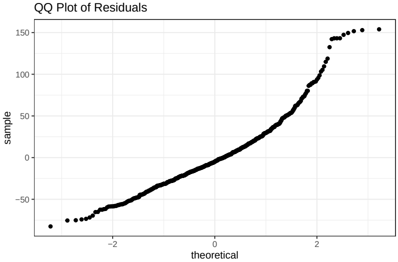
Distribution of Residuals

- We can still look at the histogram and QQ Plot of residuals, as we did for SLR:

```
ggplot(mod_res, aes(x = residual))+
  geom_histogram(bins = 30, color = "white")+
  labs(title = "Histogram of Residuals")
```




```
ggplot(mod_res, aes(sample = residual))+
  geom_point(stat = "qq")+
  labs(title = "QQ Plot of Residuals")
```



- We see some evidence that residuals are not Normally distributed (right-skew)

Section 3

Testing Model Fit

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.
    - But, when considering a model with many explanatory variables, we need to be cautious about interpreting individual p-values.

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.
  - But, when considering a model with many explanatory variables, we need to be cautious about interpreting individual p-values.
  - Suppose we had a model which contained 100 explanatory variables that were all independent of the response.
  - Approximately how many of the rows would have a p-value that is significant at the 0.05 level?

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.
  - But, when considering a model with many explanatory variables, we need to be cautious about interpreting individual p-values.
  - Suppose we had a model which contained 100 explanatory variables that were all independent of the response.
  - Approximately how many of the rows would have a p-value that is significant at the 0.05 level?
- So we may also be interested in assessing how well the overall model fits the data.

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.
  - But, when considering a model with many explanatory variables, we need to be cautious about interpreting individual p-values.
  - Suppose we had a model which contained 100 explanatory variables that were all independent of the response.
  - Approximately how many of the rows would have a p-value that is significant at the 0.05 level?
- So we may also be interested in assessing how well the overall model fits the data.
- We will test the following hypotheses:

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.
  - But, when considering a model with many explanatory variables, we need to be cautious about interpreting individual p-values.
  - Suppose we had a model which contained 100 explanatory variables that were all independent of the response.
  - Approximately how many of the rows would have a p-value that is significant at the 0.05 level?
- So we may also be interested in assessing how well the overall model fits the data.
- We will test the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad H_a : \text{ At least one } \beta_i \neq 0$$

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.
  - But, when considering a model with many explanatory variables, we need to be cautious about interpreting individual p-values.
  - Suppose we had a model which contained 100 explanatory variables that were all independent of the response.
  - Approximately how many of the rows would have a p-value that is significant at the 0.05 level?
- So we may also be interested in assessing how well the overall model fits the data.
- We will test the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad H_a : \text{ At least one } \beta_i \neq 0$$

- i.e. We will assess how likely it was to obtain slope estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ as large as those observed, if all the explanatory variables were uncorrelated with the response.

Testing For Model Fit

- Previously, we performed individual hypothesis tests to measure each explanatory variable's contribution to the model.
  - But, when considering a model with many explanatory variables, we need to be cautious about interpreting individual p-values.
  - Suppose we had a model which contained 100 explanatory variables that were all independent of the response.
  - Approximately how many of the rows would have a p-value that is significant at the 0.05 level?
- So we may also be interested in assessing how well the overall model fits the data.
- We will test the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad H_a : \text{ At least one } \beta_i \neq 0$$

- i.e. We will assess how likely it was to obtain slope estimates $\hat{\beta}_1, \dots, \hat{\beta}_k$ as large as those observed, if all the explanatory variables were uncorrelated with the response.
  - Note that $\beta_0 = 0$ is not in the null hypothesis, since we are only looking for evidence that at least one explanatory variable contributes to the response.

Partitioning Variability

- Recall that the general form of a MLR model is

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

Partitioning Variability

- Recall that the general form of a MLR model is

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

- We can split the total variability in the response variable into two pieces:
  - One representing the variability explained by the model
  - Another representing the variability given by the errors

Partitioning Variability

- Recall that the general form of a MLR model is

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

- We can split the total variability in the response variable into two pieces:
  - One representing the variability explained by the model
  - Another representing the variability given by the errors
- We write...

$$\text{SSTotal} = \text{SSModel} + \text{SSE}$$

Partitioning Variability

- Recall that the general form of a MLR model is

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

- We can split the total variability in the response variable into two pieces:
  - One representing the variability explained by the model
  - Another representing the variability given by the errors
- We write...

$$\text{SSTotal} = \text{SSModel} + \text{SSE}$$

where...

$$\text{SSTotal} = \text{Total Sum of Squares} = \sum (y_i - \bar{y})^2$$

$$\text{SSModel} = \text{Sum of Squares Explained by Model} = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \text{Sum of Squared Error} = \sum (y_i - \hat{y}_i)^2$$

Partitioning Variability

- Recall that the general form of a MLR model is

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

- We can split the total variability in the response variable into two pieces:
  - One representing the variability explained by the model
  - Another representing the variability given by the errors
- We write...

$$\text{SSTotal} = \text{SSModel} + \text{SSE}$$

  where...

$$\text{SSTotal} = \text{Total Sum of Squares} = \sum (y_i - \bar{y})^2$$

$$\text{SSModel} = \text{Sum of Squares Explained by Model} = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \text{Sum of Squared Error} = \sum (y_i - \hat{y}_i)^2$$

- Note: SSE was the quantity we sought to minimize when fitting the least squares line.

The F-Statistic

- A "good" model is one where almost all of the variability in response is due to the model, rather than the error.

The F-Statistic

- A "good" model is one where almost all of the variability in response is due to the model, rather than the error.
- Consider the following ratio:

$$F = \frac{\mathrm{SSModel}}{\mathrm{SSE}} \cdot \frac{n-k-1}{k}$$

The F-Statistic

- A "good" model is one where almost all of the variability in response is due to the model, rather than the error.
- Consider the following ratio:

$$F = \frac{\text{SSModel}}{\text{SSE}} \cdot \frac{n-k-1}{k}$$

  - Good models should have values of F (much) larger than 1
  - Models with variables uncorrelated with response should have values of F close to 1.

The F-Statistic

- A "good" model is one where almost all of the variability in response is due to the model, rather than the error.
- Consider the following ratio:

$$F = \frac{\text{SSModel}}{\text{SSE}} \cdot \frac{n-k-1}{k}$$

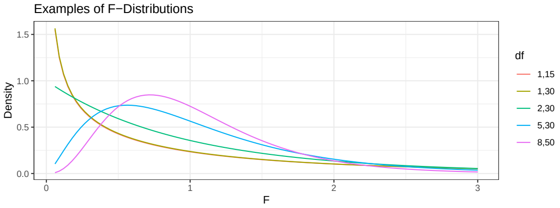
  - Good models should have values of F (much) larger than 1
  - Models with variables uncorrelated with response should have values of F close to 1.
  - The extra $\frac{n-k-1}{k}$ eliminates bias due to the number of variables and observations

The F-Statistic

- A "good" model is one where almost all of the variability in response is due to the model, rather than the error.
- Consider the following ratio:

$$F = \frac{\text{SSModel}}{\text{SSE}} \cdot \frac{n-k-1}{k}$$

  - Good models should have values of F (much) larger than 1
  - Models with variables uncorrelated with response should have values of F close to 1.
  - The extra $\frac{n-k-1}{k}$ eliminates bias due to the number of variables and observations
- Under the null hypothesis, F follows a certain theoretical distribution
  - This distribution is called the F distribution, which has two parameters: $(k, n-k-1)$



Computing P-values

- We **could** compute the value of the F statistic by hand, and then compute the p-value of the F stat using an appropriate function in R (like `pf`).

Computing P-values

- We **could** compute the value of the F statistic by hand, and then compute the p-value of the F stat using an appropriate function in R (like `pf`).
  - BUT! The `lm` function already does all of this for us.

Computing P-values

- We **could** compute the value of the F statistic by hand, and then compute the p-value of the F stat using an appropriate function in R (like `pf`).
  - BUT! The `lm` function already does all of this for us.

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_summaries(theses_mlr)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.254         0.248 1319.  36.3  36.5      43.5       0     6   772
```

Computing P-values

- We **could** compute the value of the F statistic by hand, and then compute the p-value of the F stat using an appropriate function in R (like `pf`).
    - BUT! The `lm` function already does all of this for us.

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_summaries(theses_mlr)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.254         0.248 1319.  36.3  36.5      43.5       0     6   772
```

- The `statistic` in the `get_regression_summaries` table IS the F-statistic.

Computing P-values

- We **could** compute the value of the F statistic by hand, and then compute the p-value of the F stat using an appropriate function in R (like `pf`).
  - BUT! The `lm` function already does all of this for us.

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_summaries(theses_mlr)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.254         0.248 1319.  36.3  36.5      43.5       0     6   772
```

- The `statistic` in the `get_regression_summaries` table IS the *F*-statistic.
- The `p_value` is the p-value for this statistic in the appropriate F distribution.

Computing P-values

- We **could** compute the value of the F statistic by hand, and then compute the p-value of the F stat using an appropriate function in R (like `pf`).
  - BUT! The `lm` function already does all of this for us.

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_summaries(theses_mlr)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.254         0.248 1319.  36.3  36.5      43.5       0     6   772
```

- The `statistic` in the `get_regression_summaries` table IS the *F*-statistic.
- The `p_value` is the p-value for this statistic in the appropriate F distribution.
- In the case of the Reed Theses, we see that the P-value for the F test is (very close to) 0, and so we reject the null hypothesis.

Computing P-values

- We **could** compute the value of the F statistic by hand, and then compute the p-value of the F stat using an appropriate function in R (like `pf`).
  - BUT! The `lm` function already does all of this for us.

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_summaries(theses_mlr)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.254         0.248 1319.  36.3  36.5      43.5       0     6   772
```

- The `statistic` in the `get_regression_summaries` table IS the *F*-statistic.
- The `p_value` is the p-value for this statistic in the appropriate F distribution.
- In the case of the Reed Theses, we see that the P-value for the F test is (very close to) 0, and so we reject the null hypothesis.
  - This sample gives evidence that at least one of the coefficients in the model is non-zero.

Section 4

Model Selection

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2
- Recall:

$$R^2 = \frac{\text{SSModel}}{\text{TSS}} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2
- Recall:

$$R^2 = \frac{\text{SSModel}}{\text{TSS}} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

- That is, R^2 measures the proportion of variability in the response explained by variability in explanatory variables.

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2
- Recall:
$$R^2 = \frac{\text{SSModel}}{\text{TSS}} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$
- That is, R^2 measures the proportion of variability in the response explained by variability in explanatory variables.
    - If $R^2 \approx 1$, most of the variability in response is explained by linear relationship with explanatory variables.

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2
- Recall:

$$R^2 = \frac{\text{SSModel}}{\text{TSS}} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

- That is, R^2 measures the proportion of variability in the response explained by variability in explanatory variables.
  - If $R^2 \approx 1$, most of the variability in response is explained by linear relationship with explanatory variables.
  - While if $R^2 \approx 0$, almost none of the variability is explained by the model.

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2
- Recall:

$$R^2 = \frac{\text{SSModel}}{\text{TSS}} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

- That is, R^2 measures the proportion of variability in the response explained by variability in explanatory variables.
    - If $R^2 \approx 1$, most of the variability in response is explained by linear relationship with explanatory variables.
    - While if $R^2 \approx 0$, almost none of the variability is explained by the model.
- But it turns out this formula gives a **biased** estimate of the variability in the *population* explained by the model.

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2
- Recall:

$$R^2 = \frac{\text{SSModel}}{\text{TSS}} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

- That is, R^2 measures the proportion of variability in the response explained by variability in explanatory variables.
  - If $R^2 \approx 1$, most of the variability in response is explained by linear relationship with explanatory variables.
  - While if $R^2 \approx 0$, almost none of the variability is explained by the model.
- But it turns out this formula gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we may use the adjusted R:

$$R^2_{\text{adj}} = 1 - \left(\frac{\text{SSE}}{\text{TSS}} \cdot \frac{n-1}{n-k-1} \right)$$

R^2 and Adjusted R^2

- In addition to the F statistic, we've already seen another metric from the goodness-of-fit of a multilinear model: R^2
- Recall:

$$R^2 = \frac{\text{SSModel}}{\text{TSS}} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

- That is, R^2 measures the proportion of variability in the response explained by variability in explanatory variables.
  - If $R^2 \approx 1$, most of the variability in response is explained by linear relationship with explanatory variables.
  - While if $R^2 \approx 0$, almost none of the variability is explained by the model.
- But it turns out this formula gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we may use the adjusted R:

$$R^2_{\text{adj}} = 1 - \left(\frac{\text{SSE}}{\text{TSS}} \cdot \frac{n-1}{n-k-1} \right)$$

- This adjusted R^2 is usually a bit smaller than R^2, and the difference decreases as n gets large.

Statistics Exams

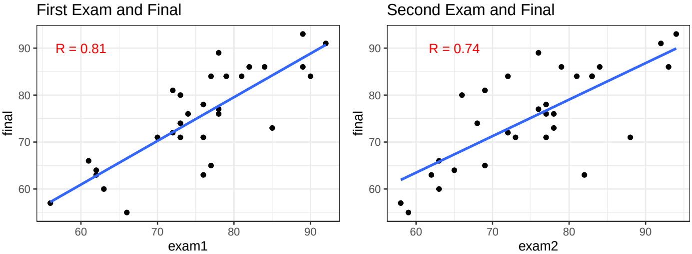
- Consider a (hypothetical) statistics class of 30 students; suppose these students have two midterm exams and a final.

Statistics Exams

- Consider a (hypothetical) statistics class of 30 students; suppose these students have two midterm exams and a final.
  - Can we use student scores on the midterms to predict their score on the final?

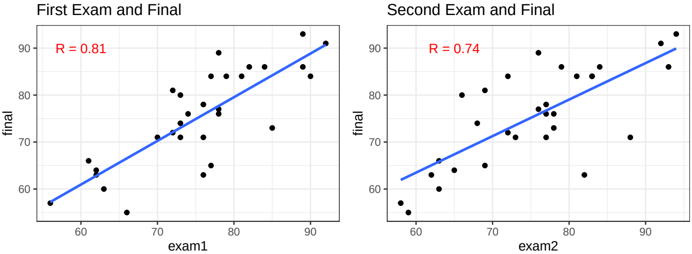
Statistics Exams

- Consider a (hypothetical) statistics class of 30 students; suppose these students have two midterm exams and a final.
  - Can we use student scores on the midterms to predict their score on the final?



Statistics Exams

- Consider a (hypothetical) statistics class of 30 students; suppose these students have two midterm exams and a final.
    - Can we use student scores on the midterms to predict their score on the final?



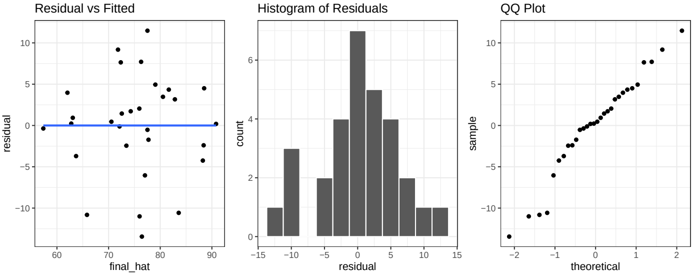
- Individually, both exams seem to have relatively strong linear relationship with the final.

Checking Conditions

- Are conditions for inference met?

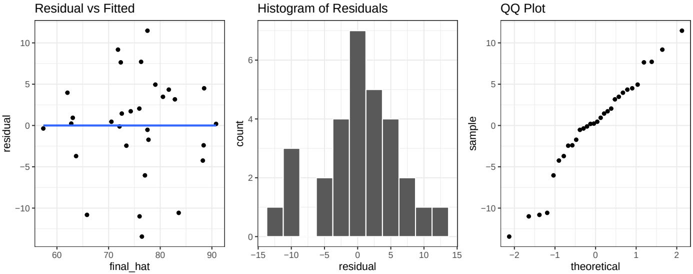
Checking Conditions

- Are conditions for inference met?



Checking Conditions

- Are conditions for inference met?



1. Linearity
2. Independence
3. Normality
4. Equal Variability

Building a MLR Model

- Let's build the model:

Building a MLR Model

- Let's build the model:

```
scores_mod <- lm(final ~ exam1 + exam2, data = stat_scores)
get_regression_table(scores_mod)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    4.830      9.76      0.49   0.625   -15.20    24.86
## 2     exam1    0.836      0.27      3.11   0.004     0.28     1.39
## 3     exam2    0.099      0.25      0.40   0.694    -0.41     0.61
```

Building a MLR Model

- Let's build the model:

```
scores_mod <- lm(final ~ exam1 + exam2, data = stat_scores)
get_regression_table(scores_mod)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    4.830      9.76      0.49   0.625   -15.20    24.86
## 2     exam1    0.836      0.27      3.11   0.004     0.28     1.39
## 3     exam2    0.099      0.25      0.40   0.694    -0.41     0.61
```

- Based on individual slope hypothesis tests. . .

Building a MLR Model

- Let's build the model:

```
scores_mod <- lm(final ~ exam1 + exam2, data = stat_scores)
get_regression_table(scores_mod)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    4.830      9.76      0.49   0.625   -15.20    24.86
## 2     exam1    0.836      0.27      3.11   0.004     0.28     1.39
## 3     exam2    0.099      0.25      0.40   0.694    -0.41     0.61
```

- Based on individual slope hypothesis tests...
  - Exam1 is a significant predictor of final score, while Exam2 is **not** a significant predictor.

Building a MLR Model

- Let's build the model:

```
scores_mod <- lm(final ~ exam1 + exam2, data = stat_scores)
get_regression_table(scores_mod)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    4.830      9.76      0.49   0.625   -15.20    24.86
## 2     exam1    0.836      0.27      3.11   0.004     0.28     1.39
## 3     exam2    0.099      0.25      0.40   0.694    -0.41     0.61
```

- Based on individual slope hypothesis tests. . .
  - Exam1 is a significant predictor of final score, while Exam2 is **not** a significant predictor.
- Hang on. . . let's build a simple linear model for `final` and `exam2`

Building a MLR Model

- Let's build the model:

```
scores_mod <- lm(final ~ exam1 + exam2, data = stat_scores)
get_regression_table(scores_mod)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    4.830      9.76      0.49   0.625   -15.20    24.86
## 2     exam1    0.836      0.27      3.11   0.004     0.28     1.39
## 3     exam2    0.099      0.25      0.40   0.694    -0.41     0.61
```

- Based on individual slope hypothesis tests. . .
  - Exam1 is a significant predictor of final score, while Exam2 is **not** a significant predictor.
- Hang on. . . let's build a simple linear model for `final` and `exam2`

```
scores_mod2 <- lm(final ~ exam2, data = stat_scores)
get_regression_table(scores_mod2)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    16.86     10.26       1.6    0.11     -4.2     37.9
## 2     exam2     0.78      0.14       5.7    0.00      0.5      1.1
```

Building a MLR Model

- Let's build the model:

```
scores_mod <- lm(final ~ exam1 + exam2, data = stat_scores)
get_regression_table(scores_mod)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    4.830      9.76      0.49   0.625   -15.20    24.86
## 2     exam1    0.836      0.27      3.11   0.004     0.28     1.39
## 3     exam2    0.099      0.25      0.40   0.694    -0.41     0.61
```

- Based on individual slope hypothesis tests. . .
  - Exam1 is a significant predictor of final score, while Exam2 is **not** a significant predictor.
- Hang on. . . let's build a simple linear model for `final` and `exam2`

```
scores_mod2 <- lm(final ~ exam2, data = stat_scores)
get_regression_table(scores_mod2)
```

```
##        term estimate std_error statistic p_value lower_ci upper_ci
## 1 intercept    16.86     10.26       1.6    0.11     -4.2     37.9
## 2     exam2     0.78      0.14       5.7    0.00      0.5      1.1
```

- In the simple mode, exam 2 **is** a significant predictor of final.

Correlated Predictors

- What is going on here? (Note that exam2 had a correlation of 0.74 with the final)

Correlated Predictors

- What is going on here? (Note that exam2 had a correlation of 0.74 with the final)
  - Recall the form of the individual slope hypotheses:

$$H_0 : \beta_{\text{exam2}} = 0, \text{ given that exam1 is in the model}$$

Correlated Predictors

- What is going on here? (Note that exam2 had a correlation of 0.74 with the final)
    - Recall the form of the individual slope hypotheses:

$$H_0 : \beta_{\text{exam2}} = 0, \text{ given that exam1 is in the model}$$

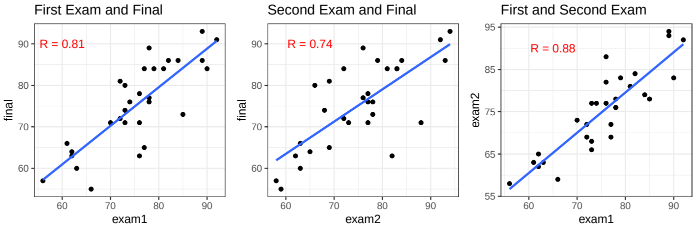
- On its own, Exam 2 is helpful for predicting the final. But IF Exam 1 is already in the model, Exam 2 is redundant.

Correlated Predictors

- What is going on here? (Note that exam2 had a correlation of 0.74 with the final)
    - Recall the form of the individual slope hypotheses:

$$H_0 : \beta_{\text{exam2}} = 0, \text{ given that exam1 is in the model}$$

- On its own, Exam 2 is helpful for predicting the final. But IF Exam 1 is already in the model, Exam 2 is redundant.
- Note that Exam 1 and Exam 2 are **highly** correlated

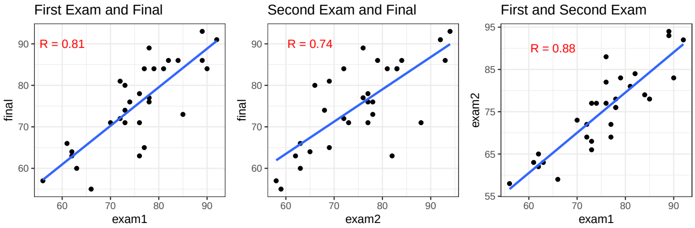


Correlated Predictors

- What is going on here? (Note that exam2 had a correlation of 0.74 with the final)
  - Recall the form of the individual slope hypotheses:

$$H_0 : \beta_{\text{exam2}} = 0, \text{ given that exam1 is in the model}$$

- On its own, Exam 2 is helpful for predicting the final. But IF Exam 1 is already in the model, Exam 2 is redundant.
- Note that Exam 1 and Exam 2 are **highly** correlated



- Once exam 1 is known, exam 2 doesn't contribute much additional information.

Model Selection

- We have 3 models for predicting final exam score:

$$\text{final} \sim \text{exam1} \quad \text{final} \sim \text{exam2} \quad \text{final} \sim \text{exam1} + \text{exam2}$$

Model Selection

- We have 3 models for predicting final exam score:

$$\text{final} \sim \text{exam1} \quad \text{final} \sim \text{exam2} \quad \text{final} \sim \text{exam1} + \text{exam2}$$

- To decide which is best, let's perform F tests and calculate R^2 and R^2_{adj}:

```
get_regression_summaries(scores_mod1)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1      0.66         0.648  35.5  5.96  6.17      54.4       0     1    30
```

```
get_regression_summaries(scores_mod2)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.541         0.525  48.0  6.93  7.17      33.0       0     1    30
```

```
get_regression_summaries(scores_mod)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.662         0.637  35.3  5.94  6.26      26.5       0     2    30
```

Summary of Model Selection

- All 3 models had statistically significant F-statistics

Summary of Model Selection

- All 3 models had statistically significant F-statistics
- Model 1 had $R^2 = 0.66$, Model 2 had $R^2 = 0.541$, and the full Model had $R^2 = 0.662$

Summary of Model Selection

- All 3 models had statistically significant F-statistics
- Model 1 had $R^2 = 0.66$, Model 2 had $R^2 = 0.541$, and the full Model had $R^2 = 0.662$
- But Model 1 had $R^2_{\text{adj}} = 0.648$, Model 2 had $R^2_{\text{adj}} = 0.525$, and the full Model had $R^2_{\text{adj}} = 0.637$

Summary of Model Selection

- All 3 models had statistically significant F-statistics
- Model 1 had $R^2 = 0.66$, Model 2 had $R^2 = 0.541$, and the full Model had $R^2 = 0.662$
- But Model 1 had $R^2_{\mathrm{adj}} = 0.648$, Model 2 had $R^2_{\mathrm{adj}} = 0.525$, and the full Model had $R^2_{\mathrm{adj}} = 0.637$
- Since Model 1 had the highest adjusted R^2 **and** was the simplest model considered, Model 1 is likely the most accurate of the 3 models.